Analysis of Video Game Sales*

Anonymous
Removed

ABSTRACT

Console gaming has been at the forefront of personal entertainment over the last few decades with an approximate, current valuation of more than $93 billion. Using the data collected from 1980 to 2016, the authors propose an in-depth analysis of the factors which drive the sales of various gaming platform. Using the Python language and it's libraries such as Panda, Numpy, and Seaborn, the authors conduct a detailed comparison of the premier platforms to identify the top contender of the console war. With illustrative proof and visualized data, PS3 is demonstrated as the most formidable platform as it consistently outperforms its peers, concerning sales, the number of hit games offered, and in the matter of maintaining consistent regional and global sales. Linear regression yields results that were not suitable for the dataset. On applying Polynomial Regression, the model achieved has a Mean Accuracy Error of 0.1849. This model fits on the sales from North America to predict sales in Europe.

KEYWORDS

Video Game, Sales, RDBMS, Data mining, Big Data

ACM Reference format:
Anonymous. 2017. Analysis of Video Game Sales. In *Proceedings of NA, NA, 2017*, 6 pages.
DOI: 10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Console wars have always been a topic of heated discussion among gamers and video game enthusiasts. With new platforms and gaming innovations such as virtual reality emerging these days, it is intriguing to delve deeper into this business and identify what factors influence this battle of high-stakes. With high throughput technologies at disposal for creating visually pleasing, jaw-dropping graphics to simulate real life and fiction based story-lines, developer companies now spend millions for a single game's birth and evolution. After going through many datasets, we had to look for something that was interesting and flexible enough to perform the various data analytics related task. Hence, the dataset publicly provided by Kirubi [8] was the one we decided to finalize.

The dataset contains 16,718 records and 16 attributes out of which 6,900 are complete with user ratings. This dataset lies within the following data domains: Entertainment, Video Games, and Sales. This dataset contains information about the sales of console-based video games during the past 37 years coupled with user ratings taken from review aggregator, Metacritic [7].

We aim to visualize the data, identify the significant trends and consequently establish relations and connections among different attributes. We have generated links and compared sales performance between game developers like Nintendo, UbiSoft, Blizzard, ActiVision etc., gaming platforms such as PlayStation, Xbox, and observed their individual market statistics. Many games have multiple iterations spanning across years or decades. Several titles are exclusive to a console which might further boost the sale of the particular platform during the release period. Some titles are released on all major platforms simultaneously. Studying such correlations, we achieved a useful hierarchy among the platforms.

The competition between the platforms and the developers can vary from being monopolistic in specific areas to being slightly stiff in others. With entertainment companies trying to consolidate their positions in the dominant regions while trying to come up with strategies to increase their hold on the emerging markets, we came up with a detailed and statistical analysis in this project. We substantiated the influence of various factors on sales of video game titles across different regions spanning 37 years and forecast the estimated growth in sales in the upcoming years. We implemented two kinds of Regression to test our models on our dataset. Linear regression yielded very unacceptable results, which implied that our data points' graph is not growing in a linear fashion. Next, we used Polynomial regression to develop a model that learns the sales from North America to predict the sales of the new games to be released in Europe.

Following is the roadmap for this paper. In Section 2, we have presented our project's design considerations. Section 3 entails the architecture in depth. Section 4 shows our steps of implementation, where we have listed down our completed tasks. Lessons learned from this project are mentioned in Section 5, and Section 6 presents how this paper is cautious about Ethical and legal matters related to our subject. Finally, Section 7 concludes our paper stating the scope of future work.

	A	B	C	D	E	F	G
1	**Name**	**Platform**	**Year_of**	**Genre**	**Publisher**	**NA_Sales**	**EU_Sales**
2	Madden NFL 2004	PS2	N/A	Sports	Electronic Arts	4.26	0.26
3	FIFA Soccer 2004	PS2	N/A	Sports	Electronic Arts	0.59	2.36
4	LEGO Batman: The Videogar	Wii	N/A	Action	Warner Bros. I	1.8	0.97
5	wwe Smackdown vs. Raw 2(	PS2	N/A	Fighting	N/A	1.57	1.02
6	Space Invaders	2600	N/A	Shooter	Atari	2.36	0.14
7	Rock Band	X360	N/A	Misc	Electronic Arts	1.93	0.33
8	Frogger's Adventures: Temp	GBA	N/A	Adventure	Konami Digital	2.15	0.18
9	LEGO Indiana Jones: The Ori	Wii	N/A	Action	LucasArts	1.51	0.61
10	Call of Duty 3	Wii	N/A	Shooter	Activision	1.17	0.84
11	Rock Band	Wii	N/A	Misc	MTV Games	1.33	0.56
12	Call of Duty: Black Ops	PC	N/A	Shooter	Activision	0.58	0.81

Figure 1: Dataset Sample (PNG).

*Produces the permission block, and copyright information

NA, NA

DOI: 10.1145/nnnnnnn.nnnnnnn

2 DESIGN CONSIDERATIONS

To ensure an efficient system of dataflow between our system components, we work with Python and connect it to the database of choice: MySQL

While considering our project design, we studied what current related work had been in this field with video game sales as the topic. The following are a few articles and papers as part of our Literature Survey.

2.1 Literature Survey

Current work on this issue, we found that in the article given by nycdatascience website [5], essential data visualization was performed with Video Game sales dataset. Pearson correlation matrix and histograms of various attributes were plotted. IBM's blog [3] published a general article about video game sales and big data's advantages in boosting sales. This article focuses on Database advancements, Mobile and Console Game analytics. No implementation has been discussed in the direction of any Data mining technique.

In the paper by Steven Emil Ehrenfeld [2] was presented at San Diego State University, where the author worked on predicting video game sales using Internet message boards discussions [2]. The dataset is entirely different than ours, but the techniques we felt were worth studying. Weka is the tool they used for their data analysis and the method used was SVM for classification and regression.

In the paper published by Jeffry Babb, and Neil Terry [1] from West Texas University, video game sales spanning 2006 to 2011 for North America are studied. The authors have presented their results of the Console wars stating which console is the best in each of the four-tier sales pyramid. They have made use of Kruskal-Wallis test. We planned our work according to the needs of the dataset. Our dataset allowed us the techniques explained in Section 4. The next section describes the architecture of our project.

3 ARCHITECTURE

The architecture of the project is as shown in Figure 2. We obtained the CSV (Comma Separated Values) dataset from the domain Kaggle [8] which provides analysts and learners with many choices of datasets. The next phase is Data Cleaning. Missing values, outliers were removed in the Data Consolidation phase. MySQL was used as the choice of RDBMS to set up the database engine. The original data were divided into normalized data tables, making sure that it follows all rules of Data Integrity and Atomicity.

Consequently, data was visualized using Python and its libraries, wherein we analyzed data correlations, causations and linkages. We plotted graphs and histograms to demonstrate various one to one and many to many relations. Ultimately, we perform the in-depth video game sales analysis and extract the insights from products of the processes mentioned above.

4 IMPLEMENTATION

4.1 Completed Tasks

4.1.1 Data Modeling. Relational Database Management System (RDBMS) uses primary keys, foreign keys, and indexes for efficient

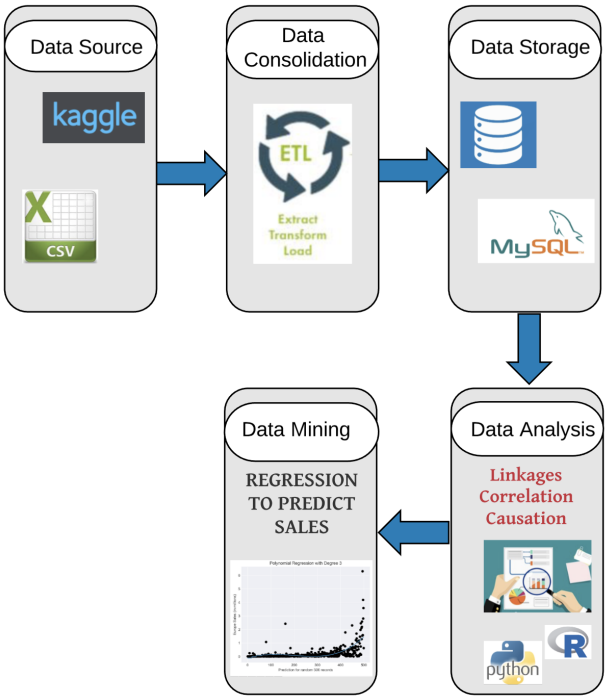

Figure 2: Architecture (PNG).

data processing. As our choice of RDBMS, we have used MySQL, which employs Structured Query Language (SQL). In the Entity-Relationship model, we aim to build tables as normalized, displayed in Figure 3. We have followed Entity Integrity Rule, which states that no primary keys should have null values. Obeying integrity constraints, we included assertions in our data tables. We have created the following tables so far:

(1) Game Description : This table contains attributes related to Game's origins such as name, year of release, platform, genre, publisher, developer etc.
(2) Sales table : This table consists of attributes about sales in various regions such as Europe, North America, Japan, apart from Global and other sales.
(3) Review table : This table contains User and Critics rating for these games along with the number of ratings received for both.
(4) Genre table : This table gives the details about the genre of the game. As this field is independent of other attributes, hence we created a separate table for it. Genre consists of only 12 unique values for more than 16000 instances. Hence, we mapped each genre to a specific number which will help with faster comparisons during the analysis.
(5) Publisher table : This table contains publisher names for all the video games.
(6) Platform table : This table contains platforms of the video games.

These tables were created as per our revised draft of the E-R diagram, as given in Figure 3. We connected our database with Python

using Pycharm IDE. We performed our analysis using Python libraries such as numpy, ggplot, matplotlib etc.

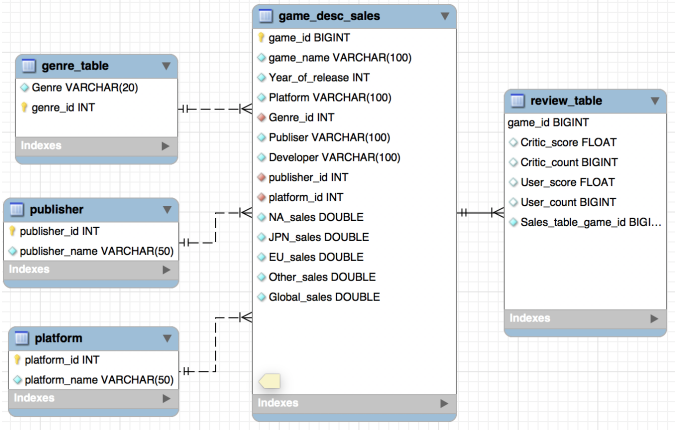

Figure 3: Entity Relationship diagram (PNG).

[H]

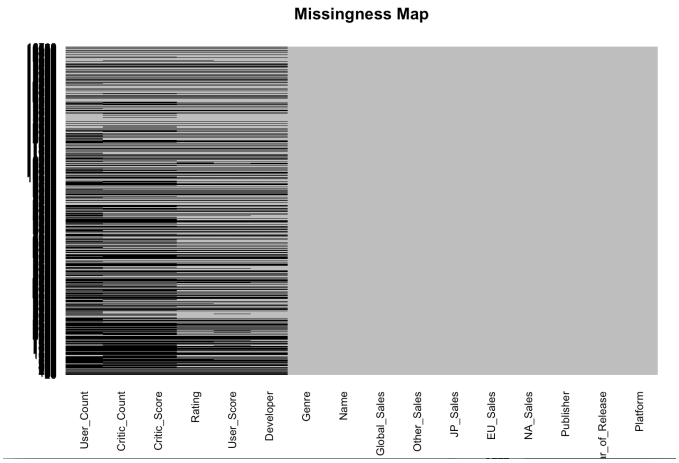

Figure 4: Missingness Map (PNG).

4.1.2 Data cleaning. As mentioned in the previous sections, on performing the exploratory analysis on raw data, we found that following attributes with huge amounts of Missing data.

(1) Critic Score - 8582 records with missing data
(2) Critic Count - 8582 records with missing data
(3) User Count - 9129 records with missing data

We plotted the Missingness Map to get an idea of how much effect deleting the records with missing data will have on overall dataset, as seen in the Figure 4.

To deal with the missing values of the critic scores and the user scores, we looked for these values on the popular gaming review domains but there is no legitimate information in one place anywhere on the web. Next, we tried finding any dataset which consists similar or subparts of such information, but were unable to get a proper result. Some sites like IGN[6] house video game ratings but it does not help our cause since they are in bits and pieces.

To conclude, we explored every possible way of finding and extracting original data but to no avail. Hence, we explored an alternative approach for this problem by establishing artificial values using the average of the score for the platform and genre or using a random generator between the permissible values but since there is a substantial number of instances with missing scores, such artificial values would highly influence the forthcoming processes. Hence, for the purpose of analysis on ratings, we shall only consider the instances with complete information and disregard the remaining data.

4.1.3 Exploratory Data Analysis. Next, we performed Pearson's Correlation and following are our observations:

(1) The resultant matrix is symmetrical and the diagonal values are completely positively correlated to themselves.
(2) Attributes User Score and Critic Score are, as expected, highly correlated.
(3) Global Sales are highly correlated with the Regional Sales across different countries.
(4) The maximum negative correlation is between User score and Year of Release. This seems fair as the year the game comes does not matter as far as its ratings are concerned.

We have then plotted various Distribution graphs between Global Sales and the various platforms and also between User scores for the various platforms. We can observe from figure 5 that PS2 has maximum lifetime sales compared to all other platforms. We can also observe Global sales of all platforms over the years in figure 6.

4.1.4 Console Wars. By studying the chart for the total sales record for every platform in the dataset, we observed that PS3, Nintendo Wii and Xbox 360 are the only platforms belonging to the same generation and having the highest yet comparable amount of sales and titles. These platforms are virtually obsolete as the next wave of consoles has taken over and hence, for determining the best console platform, we studied the performance of these three highest grossing console gaming platforms, starting by examining Figure 5. By performing an extensive sales analysis for various regions, we can conclude which platform came out as the winner. However, while determining the top gaming platform, instead of looking just at the gross sales, we also consider consistency among regions and genres, some titles vs. sales record and related factors which leads us to a more comprehensive conclusion and a clear winner.

(1) Nintendo Wii was launched in 2006 along with PlayStation 3 while Xbox 360 was launched a year before them. While comparing the global sales, it is evident that Nintendo Wii had its golden years straight from its launch till the year 2009. On the other hand, Xbox 360 had a gradual rise in the sales and the same was maintained until the next generation of Xbox was introduced in 2014. The trend was similar for the PS3 platform, but the figures fade in comparison to that of Xbox 360. Hence, attributing to the

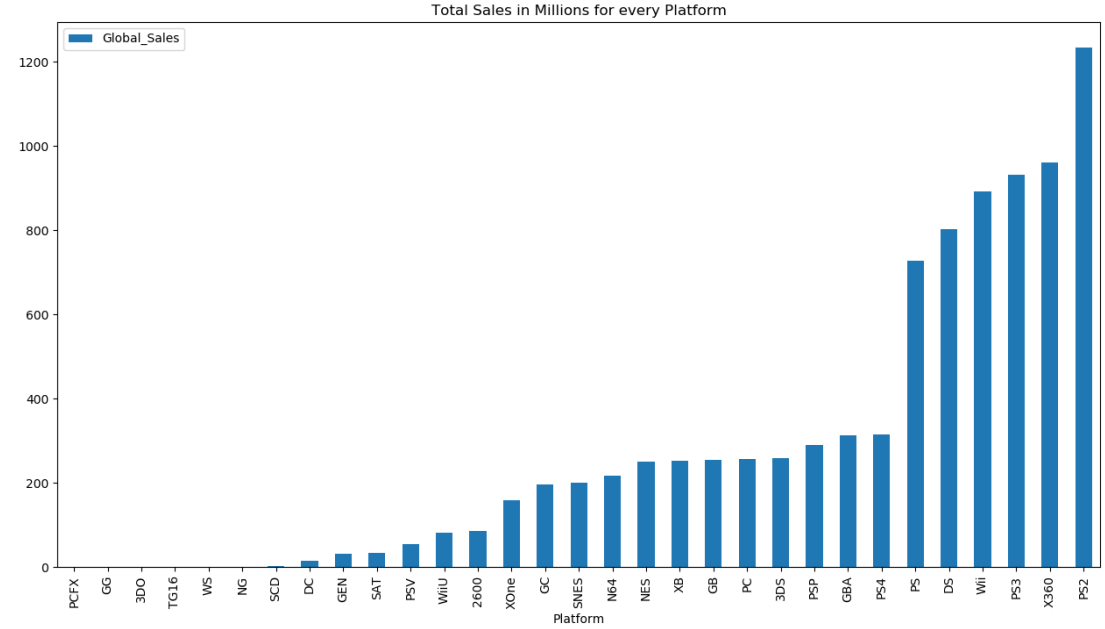

Figure 5: Total Sales vs Platform (PNG).

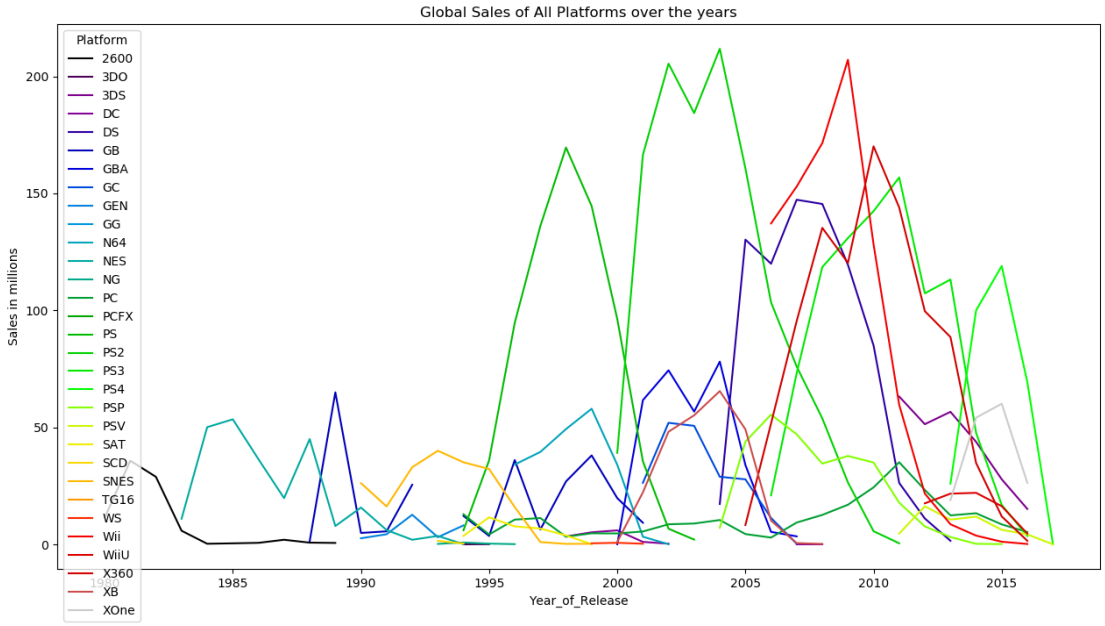

Figure 6: Global Sales per year for all platforms(PNG)

consistency of Xbox 360 over Nintendo Wii and marginally larger sales than PS3 during most years, Xbox 360 emerges a clear winner albeit by a small margin.

(2) When it comes to region-wise sales, Xbox 360 dominates the North America region mostly attesting to the fact that a USA giant, Microsoft owns the platform. When it comes to the European region, Wii dominated the sales from the outset while PS3 and Xbox 360 closed in on Wii by late 2000's and then outperformed Wii by a large margin. The contest between PS3 and Xbox 360 was a close one, but PS3 managed to beat its counterpart by a small yet considerable margin every year. Speaking of Japan, the home of Nintendo Wii, there is only one clear winner for every year with Wii at the pinnacle till 2009 and then PS3 taking the reigns till 2014. We particularly note that Xbox 360 had the most abysmal following in Japan with sales never going above 2 million in a year, a number which is comparable to that of a best selling title of the other two consoles. In the rest of the world, PS3 comes out at the top followed by Wii given its glorious sales till the end of first decade of century. Another critical observation to be made here is that Xbox 360 did not have as many sales outside its country of origin as compared to the other two platforms.

(3) While studying the sales in the context of genres, we observe that action and shooter games outperform the other ten genres by a wide margin. These two genres alone have more than 900 million copies sold to their names. Nintendo Wii, however, sold most games based on sports and miscellaneous category totaling to more than 350 million. From these observations, it is evident what no platform excels in more than a few genres, and thus, the key to a good business would be exploring the under-performing categories rather than competing to excel in the most coveted ones.

(4) Overall, PS3 maintained a good sales record spanning different regions and genres, over the years. XBox 360 has sold a few more million copies in its lifespan, but its performance outside the North American area is appalling. Wii did not have the consistency nor the numbers to outshine the others. Hence, we conclude that PS3 is the best platform of its generation.

4.1.5 Regression model. At the outset, we observed the trends of sales in different genres over the years using simple linear regression. We took into account the Mean Squared Deviation and plotted the graph along with the calculated regression coefficient. In this analysis, we found out that all but Shooter and Strategy games have a negative/steady trend. However, for plenty of genres, there is a lot of missing data for the sales. This missing data affects the slope of the trends and is not too relevant to recent sales. As a precaution, we performed the same analysis starting with the year 2000. Even this investigation didn't explain the sales precisely and hence; we implemented a different flavor of regression.

Using video games sales in North America region, we created a model to predict sales volume for the same games to be released in Europe. The key idea was to learn the sales trends from one region for video games for some makers and apply it to predict sales for the same makers' new game to be released in a different region. We implemented Polynomial regression, achieving an R2 score of 0.4091 and Mean absolute error of 0.18497. By using Sales in North America for the list of 500 games, we test to predict the sales of the same games in Europe. We used degree 3 to achieve a decently fitted model to avoid over or under fitting, as seen in Figure 8. The mistakes occurred and lessons learned from this implementation are explained under Section 5 of this paper.

5 LESSONS LEARNED

Right from selecting the dataset to the process of analysis, we learned plenty of lessons in this project. Given the freedom of choosing the dataset, our team singled out a dataset marred by missing information while carefully considering the challenge of dealing with it. However, we didn't realize how critical this aspect would play out to be in the context of the analyses. Although we could not obtain the missing data from web resources, we explored and learned about different methods to deal with the phenomenon of missing data. In retrospect, it is essential to study the structure of the data and execute a preliminary analysis of the dataset at one's disposal before initiating the process of investigation or adopt it for a comprehensive, insightful research.

At the outset, we performed Exploratory Data Analysis on the single, denormalized table in the database. However, after discovering the redundancy and duplicity of data, we undertook the process of normalization to represent the data in a more refined structure. In addition to that, we placed tight data constraints on the attributes. Our choice of data constraints was including assertions on our data tables. These preventive measures should reduce the probability of encountering anomalies in the future if more data were to be added and if the existing data were to be modified. During this project, we also learned how to determine the potential of a dataset, i.e., how it can be exploited to extract insights and identify the questions which can be answered through different analysis techniques.

For implementing Regression, linear regression did not give a fitting model to our dataset. We couldn't correctly outline the trend properly. On using polynomial regression with degree of polynomial as 1, or 10 yielded a model that was under-fitted and over-fitted respectively. With degree 10 the model curved a lot along the data points thus making it a tight fit. Hence, to make the model for accepting the values in a natural fit, we decided on finalizing our degree of polynomial to be 3.

6 ETHICAL AND LEGAL ISSUES

We tracked the origin of data to the website VGChartz[4], which collects data about game sales from different web resources. The website possesses proprietary hold over the information since the data is sourced using proprietary tools. We have referenced them in our paper as per the terms of use of the website in question. However, the nature of the information does not entail any ethical issues if we publish the findings of this project. Moreover, the collection of such information does not bypass personal or organizational privacy given the fact that historical game sales data is widely available to be extracted from different resources around the web and studies on such data have been referenced in the literature review.

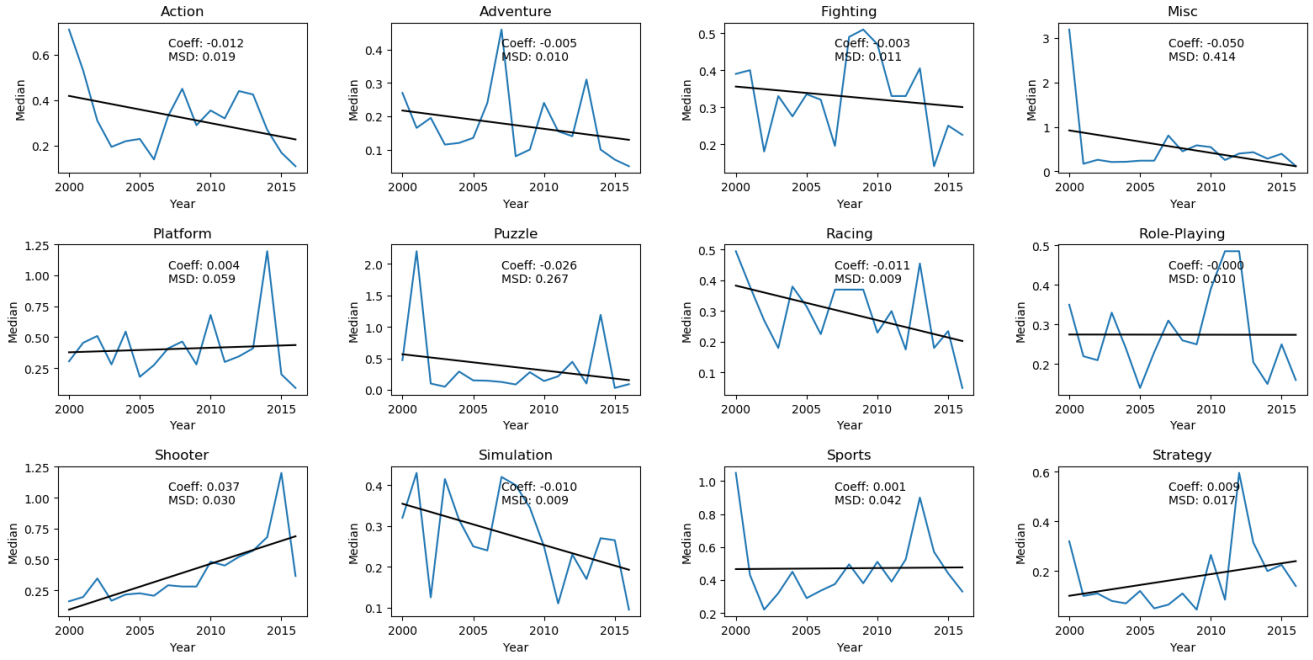

Figure 7: Linear Regression Model (PNG).

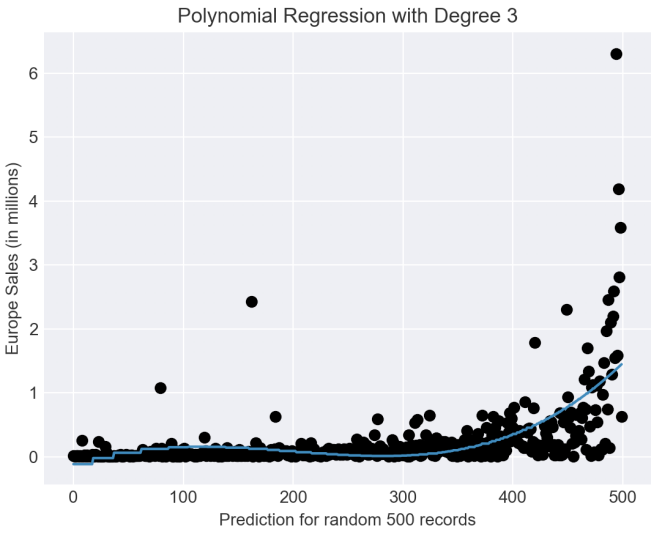

Figure 8: Polynomial Regression Model (PNG).

7 CURRENT STATUS & FUTURE WORK

The results of this paper aid us in understanding what factors drive and impede the growth of individual gaming platforms. We substantiated the performance of the best-selling platforms with extensive illustration, but considering the lack of information about the development costs, our results cannot determine the profitability of these businesses. Using regression, we were able to develop a model that fits the Dataset decently with a Mean accuracy error of 0.1849 where the best value of MAE is 0.0. However, future work would be testing the model even further with new unseen records. Classification based on ratings can also be implemented. The next steps entail discovering how Microsoft can perform better outside the home market and ascertain how the platforms would fare without their hit franchises. With the information on the current year sales, the combined analysis of the top three platforms across different generations can shed light on more insightful information and conclude an absolute and clear winner of the console war.

REFERENCES

[1] 2011. Comparing Video Games Sales by Platform. wtamu.com. (March 2011). http://swer.wtamu.edu/sites/default/files/Data/303-1102-1-PB_0.pdf.
[2] 2011. Predicting video game sales using analysis of Internet Message Board Discussions. calstate.com. (March 2011). http://sdsu-dspace.calstate.edu/bitstream/handle/10211.10/1073/Ehrenfeld_Steven.pdf?sequence=1.
[3] 2012. Big data analytics for video, mobile, and social game monetization. ibm.com. (July 2012). https://www.ibm.com/developerworks/library/ba-big-data-gaming/index.html.
[4] 2016. VGChartz game database. vgchartz.com/gamedb/. (December 2016). http://www.vgchartz.com.
[5] 2017. Analysis of Console Game Sales across Regions and Genres. nycdatascience.com. (February 2017). https://nycdatascience.com/blog/student-works/r-shiny/analysis-console-game-sales-across-regionsgenres/.
[6] 2017. IGN Game reviews. (October 2017). http://www.ign.com/reviews/games.
[7] 2017. MetaCritic Game reviews. ign.com. (October 2017). http://www.metacritic.com/game.
[8] Rush Kirobi. 2017. Video Game Sales with Ratings. Kaggle.com. (January 2017). https://www.kaggle.com/rush4ratio/video-game-sales-with-ratings.